Turning Medical Data into Decision-Support Knowledge

[1]Bohren, Benjamin F. and [1,2]Hadzikadic, Mirsad
[1]Carolinas Medical Center and [2]UNC-Charlotte

Advances in information collection and analysis are reaching the point of providing physicians with the help of computer-based assistants. These systems will provide rapid second opinions to physicians in a clinical setting as well as assist them in the analysis of large sets of patient descriptions for research purposes. This paper presents INC2.5 as such a decision-support system. INC2.5 extracts information from databases of previously seen patients to build a decision tree which is used to predict the outcome of new patients on a chosen variable. The concept of matching new patients with the most similar previously seen patient, on which INC2.5 is based, can be easily understood by its users. Further adding to INC2.5's ease of use is its flexibility in allowing users to customize decision trees to their liking. In order to convey the uncertainty of the environment, INC2.5 presents all decisions with a confidence factor.

INTRODUCTION

Doctors are faced with the process of making critical decisions everyday. While years of experience can help them be more efficient and accurate in their diagnoses, automated decision support tools can add an extra measure of confidence. However, their usefulness will extend only as far as the user's trust. This trust can be achieved by providing accurate results over an extended period of time. The clinical advantage of such systems will extend well beyond helping one doctor make better decisions. Eventually they will facilitate the spread of knowledge to institutions with fewer resources.

For the physician involved in research, a decision support system should be able to assist him/her in finding correlation's amongst large quantities of data. Additionally, the system should be able to decipher which predictor variables appear to be most relevant to the selected predictive variable. With such information, the physician could minimize the number of tests required for an accurate diagnosis, thereby reducing the cost of care.

INC2.5

INC2.5[6] is a general classification system capable of uncovering patterns of relationship amongst records stored in databases. INC2.5, similarly to COBWEB[1], CLASSIT[2], CYRUS[3], UNIMEM[4], and CLUSTER/2[5], works in an incremental manner, incorporating new knowledge one experience at a time. This is similar to the way humans learn over time and can be viewed as analogous to the physicians pattern of seeing one patient at a time. Hereafter, the term *patient* will be used to replace experience when referring to a single encounter.

INC2.5 differs from other classification systems in several key issues: (a) evaluation function, (b) tree-building operators, and (c) classification and prediction algorithms. First, INC2.5 uses a similarity-based patient evaluation function which optimizes patient's predictive variable only with respect to the *most similar* group of previously seen patients rather than with respect to all available patients. Second, INC2.5 uses unique tree-building operators, pull-in and pull-out, to reverse unwarranted decisions made Early in the classification process when less information was available. Finally, the classification and prediction algorithms are designed to maximize predictive performance of the system in the presence of noise.

Classification

INC2.5 uses the evaluation function to classify and group patients based on similarities and dissimilarities found in their patient descriptions[6,7]. Each patient description contains a list of information the physician deems possibly relevant to the diagnosis in question. This list contains information ranging from patient age to specific test results and hereafter will be referred to as the patient's *variables*. Each group of patients within the decision tree, will be referred to as a *category*.

0195-4210/94/$5.00

The process of building the decision tree is known as *classification.* Each new patient is classified into the branch of the tree which maximizes the *evaluation function.* The evaluation function can be broken into two components, similarity and cohesiveness. Similarity is used for both classifying previous patients and predicting the class membership of new patients. Cohesiveness calculates the average similarity of all pairs of patients contained in a category.

The similarity of two patients is based on the comparison between the two sets of variables. The function is derived from the contrast model[8], which defines similarity as a linear combination of common and distinctive variables. The following equations review the similarity function **s(A,B)** where **A** and **B** denote descriptions of patients or categories **a** and **b**, respectively; **c(A,B)** represents the contribution of **a**'s and **b**'s common variables; and **d(A,B)** introduces the influence of the variables of **a** not shared by **b**.

$$s(A,B) = \frac{sim(A,B) + sim(B,A)}{2}$$

$$sim(A,B) = \frac{c(A,B) - d(A,B)}{c(A,B) + d(A,B)}$$

The degree of similarity between **A** and **B** ranges from -1.0 to +1.0. In an extreme case where **A** and **B** are identical then **c(A,B)** would have a value, while **d(A,B)** and **d(B,A)** would equal zero, yielding the similarity measure equal to 1.0. Conversely, if they are completely dissimilar all the values will be in **d(A,B)** and **d(B,A)** while **c(A,B)** equals zero, yielding -1.0.

The cohesiveness measures also ranges from -1.0 to +1.0 and reflects the similarity between all member patients within a given category. A category will have a cohesiveness measure of 1.0 if and only if all member patients are identical. On the other hand, a category in which member patients are completely opposite would have a cohesiveness measure of -1.0.

Prediction

Prediction follows classification and works on the same principle. INC2.5's goal is to maximize the similarity score between the patient in question and one of the patients in the decision tree. This is efficiently achieved by only searching the branches which maximize similarity, thus providing an effective method of indexing decision trees. Once the system has searched the appropriate branches, the outcome of the most similar patient will be used as the prediction for the new patient. Furthermore, the system will report the similarity score which is in essence the degree of confidence INC2.5 has in the prediction.

Customization of Decision Trees

INC2.5 is flexible enough to allow physicians to build custom decision trees. The easiest way to customize a decision tree is to build it using a selected group of patients, patient variables, or both. For example, a physician may wish to have only his/her patients included in the tree. While this might limit the variety of patients, it could ensure that all data was collected more consistently.

Besides changing the input data, INC2.5 allows the user to adjust *certainty* (CT) and *variable* (VT) thresholds which can be used to fine tune a decision tree to be most effective within a given domain. The certainty threshold is used to determine if there is enough evidence to support a prediction. In order for INC2.5 to make a prediction, the similarity score of the closest match must exceed the certainty threshold. With the default value, CT = -1, INC2.5 will always make a prediction. At higher degrees of certainty, it is possible that no patient will be found with a similarity greater than the required certainty. In this case any number of methods could be used to provide the user feedback including the use of prior probability or stating that no prediction is possible with the given information.

The variable threshold requires a greater understanding of INC2.5. This threshold attempts to weed out patient variables which are inconsistent with those of other patients within the same category. Inconsistent values can occur either in variables which are less relevant to the diagnosis or via data entry errors. When comparing a patient with an existing category, the category will have multiple values for each variable representing the union of all member patient's values for the variable in question. The patient is said to have this variable in common if the category has the same value for the variable. For example, assume a category consists of eight red and two blue members. In this situation if

VT = 0, the default value, all objects with either red or blue color would have the color variable in common. On the other hand, if the threshold value is greater than the percentage of blue objects, VT > 0.2, any new object with the color blue will NOT have the color variable in common with this category. Consequently, the higher the incidence of a variable in a given category, the higher its relevance to the category description.

In general, INC2.5 is a flexible decision support system whose results can be used in a clinical setting as a second opinion or in a research setting for data analysis.

TEST DATABASE

The medical database used for testing INC2.5 include breast cancer, general trauma, and low back pain. While INC2.5 performed consistently across all domains, we will restrict our discussion to the widely available breast cancer database[1] which was retrieved from the machine learning repository at University of California at Irvine. It consists of 699 patients with two ideal classes, YES and NO. YES means the patient had a recurrence of breast cancer within five years, and NO means there was no recurrence during the five year period. Within the database, 458 of the patients are benign, NO, and 241 are malignant, YES. For each patient there are nine associated variables.

TESTING METHODOLOGY

INC2.5 results, presented in the following section, show performance for various tree sizes and threshold settings. Each point in the graphs represents an averaged performance over a series of ten runs. For each run, a classification set and prediction set of patients were randomly selected from the database so that no patient appeared more than once in the union of the two sets. The classification set was then used to build the decision tree subsequently utilized to predict the outcome variable, YES/NO reoccurrence, for each patient in the prediction set.

The outcome variable is only used once a match has been found, at which time INC2.5 predicts the same outcome for the new patient. In other words, the outcome variable neither influences the classification process nor guides the prediction process.

PERFORMANCE ANALYSIS

This section will perform a step by step analysis of the database. The steps presented here are just a guideline to follow. Tests can be performed in any order once a user is familiar with INC2.5.

Initial Learning Curve

Step one is to build a learning curve using default values for both thresholds. A learning curve will answer two important questions: (1) Is INC2.5 predicting significantly better than random guessing, i.e. 50% for two category domains? (2) What is the optimal tree size required to maximize accuracy while minimizing time? As demonstrated in Figure 1, the initial learning curve (CT = -1) is performing significantly better than random guessing.

When comparing the tree size to its performance the learning curve grows as expected by gradually improving with the size of the tree. It flattens just past the tree size of 100 patients. This would indicate that a random sample of 100 patients is sufficient to distinguish amongst the various outcomes. In other words, patient samples greater than 100 added no additional knowledge to the decision tree.

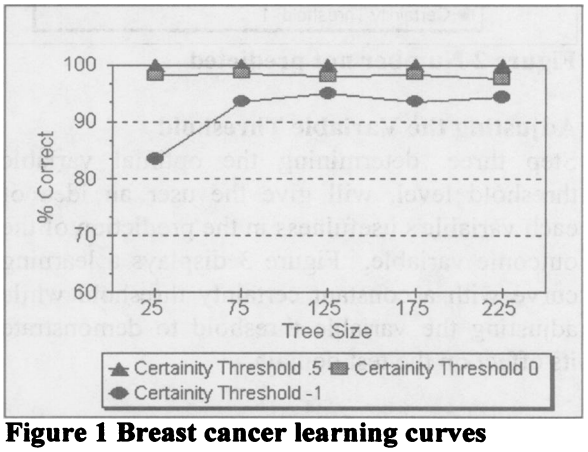

Figure 1 Breast cancer learning curves

Adjusting Certainty Threshold

Step two is to adjust the certainty threshold (CT), thereby increasing the desired level of confidence for each prediction. Figure 1 shows that when CT = 0, meaning at least half of the variables are identical, the performance of INC2.5 is consistently better than that obtained

[1]The breast cancer database was obtained from Dr. William H. Wolberg at the University of Wisconsin Hospitals, Madison.

with CT = -1. Increasing the threshold to 0.5 improves performance to almost flawless prediction.

On the down side, Figure 2 shows the number of patient diagnoses INC2.5 was unsure about. At the original threshold, CT = -1, INC2.5 will always give a prediction, thus the flat curve at zero patients not predicted. Moving up to a zero threshold, the domain shows a curve with a negative slope indicating, as one would expect, with a larger classification set INC2.5 is able to find good matches for more patients in the prediction set. The outcome remains consistent with expectations at the 0.5 threshold level as well.

By evaluating Figures 1 and 2 we can conclude that INC2.5 will perform well in both small and large data sets, but with a degree of confidence directly proportional to the tree size.

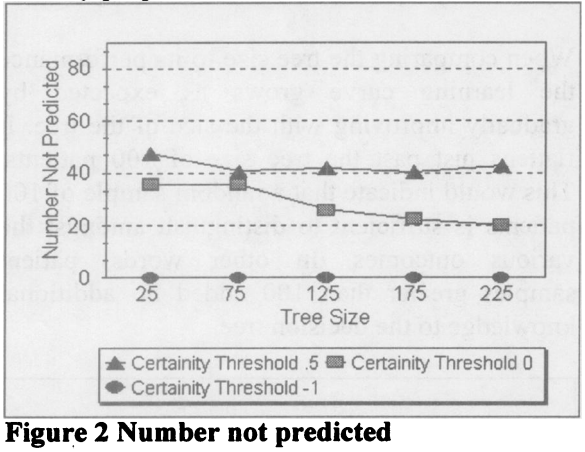

Figure 2 Number not predicted

Adjusting the Variable Threshold

Step three, determining the optimal variable threshold level, will give the user an idea of each variable's usefulness in the prediction of the outcome variable. Figure 3 displays a learning curve with a constant certainty threshold while adjusting the variable threshold to demonstrate its effect on the test domain.

The variable threshold has the most effect on databases with little consistency amongst variable values. In these cases, a high threshold would eliminate many variables thereby narrowing information used to make decisions.

When set properly, the threshold will eliminate the infrequent variable values which cloud the decision process, thereby maximizing the ability to find good matches in the tree. For example, Figure 3 demonstrates how the threshold worked effectively for the tree of twenty-five patients improving performance as the system eliminated a greater number of variables.

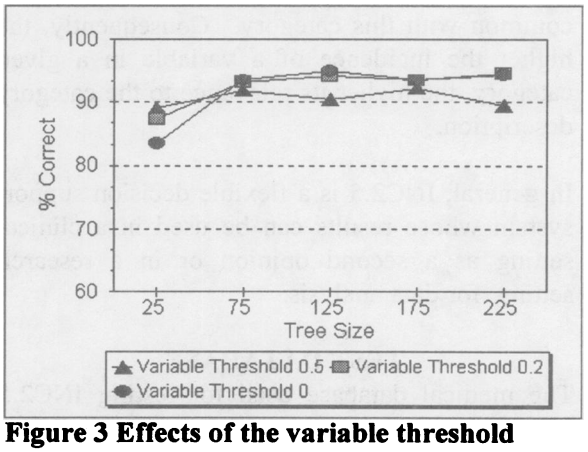

Figure 3 Effects of the variable threshold

Notice for larger patient sets the highest threshold caused a drop in performance. This indicates that too many variables were eliminated from the decision process. The 0.2 threshold neither improved nor hurt performance on the larger patient sets while helping the decision process in the smallest patient set. Therefore, for the breast cancer database 0.2 would be the optimal threshold.

Determining Variable Relevance

Step four, attempting to reduce the number of variables required for the decision process can produce two major benefits: (1) a reduced set of attributes which would make statistical analysis easier and (2) if proven to be effective, many costly tests could be eliminated during the diagnosis process. The variable reduction process assumes INC2.5 has been successful at forming a tree with categories containing a majority of patients having the same outcome. In that case, the variables common to most patients are good predictors of the outcome, while variables which vary widely over the category are not relevant to the outcome.

For this test a tree of one-hundred patients was built using all provided variables. Then INC2.5 systematically reduced the number of relevant variables by incrementing the relevancy measure. The relevancy measure is an experiential indicator sensitive to the depth of the tree as well as the frequency and

distributions of variable values. Figure 4 shows prediction results plotted against the number of variables used during training and prediction. The same training and prediction sets were used for all runs thereby making the variables used the only varying factor.

During this test, the relevancy measure is incremented one percent at a time and new prediction results are generated at each percentile where one or more variables drop out.

Figure 4 shows that the number of variables can often be dramatically reduced without significant loss in performance. In this example, the prediction rate was maintained while the number of variables dropped from nine to one, that being uniformity of cell size.

The benefits of reducing the amount of information required for accurate diagnosis will be reaped in both research and clinical terms. Researchers will save precious resources during future data collection. Clinical staffs will save time and money by eliminating unnecessary tests which in turn could have a positive effect on controlling the cost of health care.

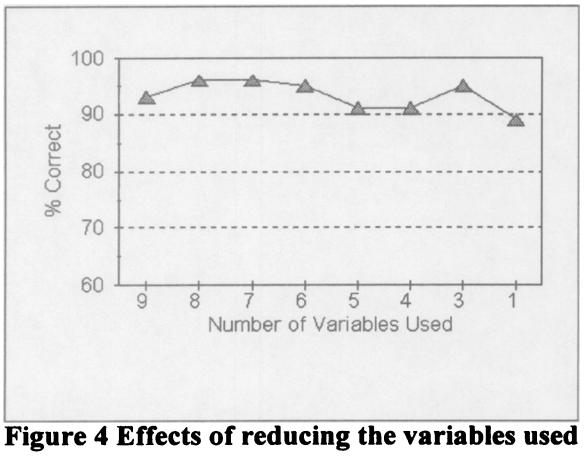

Figure 4 Effects of reducing the variables used during classification and prediction.

CONCLUSION

INC2.5 has the ability to look at sets of patients and provide quality information concerning the predictability of an outcome variable as well as the relevance of patient variables with respect to the outcome. Once the user has formed general characteristics of the data, he/she can fine tune INC2.5 for optimal performance. Increased performance can be achieved via three basic methods:

1. adjusting certainty threshold
2. adjusting variable threshold
3. eliminating irrelevant variables

The certainty threshold is easiest to use and generally yields the greatest improvement in accuracy. The variable threshold requires a greater understanding of the analysis process, thus making it more difficult to find an optimal value, but it adds to the accuracy of the prediction results. Finally, eliminating variables can preserve the quality of the output while reducing the cost of data acquisition.

Future work will concentrate on providing the ability to evaluate linear variables and enabling INC2.5 to read XBase data storage format.

ACKNOWLEDGMENT

The authors would like to thank Amber Harrington for her constant support and encouragement as well as her proof-reading of this manuscript.

Reference

[1] Fisher, D. H. Knowledge Acquisition Via Incremental Conceptual Clustering. In *Machine Learning*, **2**, 2 (1987) 139-172.

[2] Gennari, J. H., Langley, P., and Fisher, D. H. Models of Incremental Concept Formation. In *Artificial Intelligence*, **40**, 1-3(1989) 11-59.

[3] Kolodner, J. L. Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model, Lawrence Erlbaum Associated, Publishers, London, 1984.

[4] Lebowitz, M. Experiments with Incremental Concept Formation: UNIMEM. In *Machine Learning*, **2**, 2 (1987) 103-138.

[5] Michalski, R.S., and Stepp, R.E. Learning From Observation: Conceptual Clustering. In Machine Learning: An Artificial Intelligence Approach, R.S. Michalski, J.G. Carbonell, and T. M. Mitchell (eds.), Morgan Kaufman Publishes, Inc., Lao Altos, CA, 1983.

[6] Hadzikadic, M., Automated Design of Diagnostic Systems. *Artificial Intelligence in Medicine Journal*, **4** (1992a) 329-342.

[7] Hadzikadic, M. and Bohren, B. F., INC2.5: A Concept Formation System. Technical Report 008-1994, UNCC, 1994.

[8] Tversky, A. Features of Similarity. *Psychological Review*, **84** (1977) 327-352.